Biotechnology Division - Table of Contents

Division Overview

The Biotechnology Division is the focus of the NIST effort addressing critical measurement and data needs for the rapidly developing biotechnology industry.

MISSION

The mission of the Division is to provide measurement infrastructure necessary to advance the commercialization and application of biotechnology. This is achieved by developing a scientific and engineering technical base along with reliable measurement techniques and data to enable U.S. industry to produce biochemical products, and to enable the government to apply advances in biotechnology to the benefit of societal needs. The Division has established a variety of long-range research projects to maintain critical expertise needed for the development of Standard Reference Materials, Standard Reference Databases, and advanced measurement methods. The Division fosters collaboration among NIST scientists conducting biology-related research, and raises the visibility of the NIST and Chemical Science and Technology Laboratory Programmatic Areas that have strong biological focus.

Division scientists participate in scientific meetings, topical workshops, and numerous national and international organizations such as: Biotechnology Industry Organization (BIO), IUPAC Commission on Biophysical Chemistry, ASTM Committee E-48 on Biotechnology, the International Measurement Standards Consultative Committee for the Amount of Substance (CCQM), Bioanalytical Working Group. Division members were also active as reviewers for the NIST Advanced Technology Program (ATP), for several NSF and DOE programs, NIH study section panels, and for the Office of Science and Technology Policy on issues related to bioterrorism defense, and with the Department of Justice on issues related to forensics and human identification.

The staff of the Biotechnology Division consists of 50 NIST employees and a comparable number of contract researchers, guest scientists, and post-doctoral fellows. The Division is organized into four groups: (1) **DNA Technologies**; (2) **Bioprocess Engineering**; (3) **Structural Biology**; and (4) **Bimolecular Materials**. Brief descriptions of technical highlights from each Group are given below.

Selected Program Highlights:

DNA TECHNOLOGIES

The DNA Technologies Group has research projects that are included in the Program Areas of Health and Medical Products, Forensics and Homeland Security, and Food and Nutritional Products. **Advanced mass spectrometry measurements of DNA damage** are used to describe the cellular accumulation of two major oxidative stress-induced DNA lesions in cells of Cockayne syndrome (CS) patients after exposure to ionizing radiation. As a disease with implications for understanding the human aging process, these studies are undertaken as a collaborative effort with scientists at the National Institute of Aging. Projects in the area of **DNA diagnostics for the detection of human**

disease include the NIST-National Cancer Institute Biomarkers Validation Laboratory (BVL), the NIST component of the Early Detection Research Network which serves to refine recently discovered cancer biomarkers, and to format new research tests for field trials in EDRN clinical laboratories. Another area is the study of cellular biomarkers that can be used for quality assurance of tissue-engineered medical products in terms of genetic damage. In the **human identity/forensic science** project, the group focuses on new methods for DNA profiling, ranging from developing well-characterized DNA standards for restriction fragment length polymorph-isms (RFLPs) to performing research for rapid determination of DNA profiles by polymerase chain reaction (PCR) amplification and automated detection of fragments. New methods were developed for identification of victims of the World Trade Center (WTC) disaster of September 11, 2001 where the high degree of DNA fragmentation due to the severe environmental conditions has meant that only about 50% of the specimens yielded results with standard DNA testing methods.

BIOPROCESS ENGINEERING

The Bioprocess Engineering Group (**http://www.cstl.nist.gov/div831/bioprocess/**) is concerned with the development of measurement methods, databases, and generic technologies related to the use of biomolecules and biomaterials. The results are directed at the biomanufacturing and pharmaceutical industries and, most recently, to Homeland Defense, where there are needs for the detection and quantification of very small amounts of biological materials. The effort is organized into four project areas that are part of the Pharmaceuticals and Biomanufacturing, Food and Nutritional Products, and Forensics and Homeland Security Programs. In the **biospectroscopy** project, one study is directed at investigation of the mechanism of fluorescence resonance energy transfer (FRET) when it is used to quantify the extent of a polymerase chain reaction (PCR). In the figure, FRET efficiency is seen to decline by five-fold as a function of fluorophore separation, counted as number of nucleic bases between the fluorophore labeling sites. In the **biocatalysis** project, enzyme characterization is being carried out to address industrially important biotransformation problems such as those found in hydroxylation and aromatic amino acid metabolic pathways. The methods used include site-directed mutagenesis, circular dichroism, ellipsometry, spectroelectro-chemistry, and X-ray diffraction to characterize several key steps along metabolic pathways. In the **biothermodynamics** project, chromatography and microcalorimetry are used with chemical equilibrium analysis of complex reacting systems to develop thermodynamic data for industrially important biotransformations that are included in the NIST Standard Reference Database "Thermodynamics of Enzyme-catalyzed Reactions." A new project, **bioterrorism research**, has recently started to develop standard methods, materials and data related to the national efforts to defend against threats of biological warfare.

STRUCTURAL BIOLOGY

The Structural Biology Group at the Center for Advanced Research in Biotechnology (CARB) is focused in key areas of industrial biotechnology, especially in the Pharmaceuticals and Biomanufacturing Program. These areas are supported at CARB through a highly interactive group of scientists, from both the University of Maryland Biotechnology Institute (UMBI) and NIST. In the project, macromolecular structure

determination by X-ray crystallography, a new effort has been launched in structural genomics. The goal is to develop high-throughput approaches for elucidating the structures and functions of all the proteins encoded by entire genomes, with a focus on determination of the structures for 'hypothetical' proteins of microbial genomes that may be useful drug targets. The molecular structure and dynamics project includes a study of the dimerization of two homologous strands of genomic RNA, an essential reaction in the replication of retroviruses such as HIV-1 (see figure). Results from the physical, molecular and cellular biochemistry project on the structures and interactions of key recognition elements in G coupled protein receptors suggest new, quantitative models for signal transduction pathways in vision and viral infection. The energetics of enzyme-catalyzed reactions are being studied by differential stopped flow microcalorimetry. The temperature dependence of the kinetics of the acylase hydrolysis reaction has recently been determined. The staff of the computational biology project has been investigating the reaction mechanisms of two enzymes, zinc lactamase and chorismate mutase, that are representative of their enzyme class. The bioinformatics project is striving to establish data uniformity and to develop the physical archive of the NSF/DOE/NIH supported Protein Data Bank (**http://pdb.nist.gov/**) within the Research Collaboratory for Structural Bioinformatics (RCSB) partnership that includes groups from Rutgers University, the University of California San Diego Supercomputer Center, and the University of Wisconsin.

BIOMOLECULAR MATERIALS

The Biomolecular Materials Group studies the behavior of biological molecules and adapts them for novel technological and scientific applications, and to emerging needs of bioterrorism research. Measurement methods including surface plasmon resonance, IR spectroscopy, ellipsometry, electrophysiology, impedance spectroscopy, chemical synthesis, atomic force microscopy, and confocal microscopy are combined with computer simulations to carry out projects in nanobiotechnology, tissue engineering, and mitochondrial proteomics. In the nanobiotechnology project, single nanometer-scale pores were employed to study biological transport processes, to read information within single biomolecules and to detect multiple analytes in solution. In the tissue engineering project, the need for biomarkers, physical standards and measurement technologies for tissue engineering are being addressed to assure quality control during manufacturing and storage of engineered medical products. A method for reproducibly and reliably fabricating films of collagen to provide surfaces on which cells can be grown is under development. Data shown in the figure illustrate that the films can induce morphological changes in vascular smooth muscle cells that mimic their response in healthy and diseased arteries. The ability to characterize these films with surface analytical techniques permits the evaluation of how changes in the collagen substrate influence cellular responses, potentially leading to reference materials. In another study, the ability of chemokines to interact with G-proteins, which are important target molecules of the pharmaceutical industry, is interrogated with surface plasmon resonance spectroscopy. Cell fragments containing the transmembrane proteins are immobilized to a solid support where cells containing the G-protein CCR5 bind to a surface. An antibody against CCR5 subsequently binds to these membranes, but not to a control surface. These results suggest that the technique may be useful for detecting the binding of small ligands to

these receptors. The emphasis of the mitochondrial proteomics project will be to address needs of the mitochondrial and proteomics communities as outlined in a September 2002 workshop and to develop general protocols for handling and characterizing membrane associated proteins.

The Research Collaboratory for Structural Bioinformatics - Protein Data Bank:
http://www.rcsb.org/
http://rcsb.nist.gov/

The Biological Macromolecule Crystallization Database:
http://wwwbmcd.nist.gov:8080/bmcd/bmcd.html

The Short Tandem Repeat DNA Internet Database:
http://www.cstl.nist.gov/biotech/strbase/

Thermodynamics of Enzyme-Catalyzed Reactions:
http://wwwbmcd.nist.gov:8080/enzyme/enzyme.html

HIV Protease Database:
http://srdata.nist.gov/hivdb/

Biomarkers of Oxidative DNA Damage used to Detect Genetic Changes in Tissue Engineered Skin

CSTL Program: Biomaterials

Authors: *H. Rodriguez, P. Jaruga, M. Birincioglu, P.E. Barker, C. O'Connell, and M. Dizdar*

Abstract: Scientific studies have shown that free radicals are culprits of many diseases of the older population. The most common biomarker for assessing free radical-induced oxidative stress in living cells is oxidative damage to DNA. Recently, liquid chromatography/tandem mass spectrometry (LC/MS/MS) and liquid chromatography/mass spectrometry (LC/MS) emerged as new techniques for the measurement of modified nucleosides in DNA. Using gas chromatography/mass spectrometry (GC/MS), a total of 5 genomic DNA biomarkers in tissue-engineered skin (TestSkin II®, Organogenesis, Inc) was screened and the levels of damage compared to control cells, including neonatal human dermal fibroblasts, neonatal human epidermal keratinocytes, cultured HeLa cells, and commercially available calf thymus DNA. Results showed that the level of oxidative DNA damage was found to be at background/endogenous levels (approximately 1-10 modified molecules/10^6 DNA bases).

Purpose: Because scientific studies have shown that free radicals are culprits of many diseases of the older population, it is important to determine if any significant level of free radical-induced damage to DNA (a biomarker of cellular inflammation) has occurred to the tissue-engineered product. The most common biomarker for assessing free radical-induced oxidative stress in living cells is oxidative damage to DNA. We have developed liquid chromatography/tandem mass spectrometry (LC/MS/MS) and liquid chromatography/mass spectrometry (LC/MS) as new techniques for the measurement of modified nucleosides in DNA.

Major Accomplishments: Using gas chromatography/mass spectrometry (GC/MS), a total of 5 genomic DNA biomarkers in tissue-engineered skin (TestSkin II®, Organogenesis, Inc) was screened and the levels of damage compared to control cells, including neonatal human dermal fibroblasts, neonatal human epidermal keratinocytes, cultured HeLa cells, and commercially available calf thymus DNA. The biomarkers consisted of FapyAdenine, FapyGuanine, 8-OH-Guanine, 5-OH-Uracil, and 5-OH-Cytosine. For 8-OH-Guanine (a free base), its nucleoside form (8-OH-dGuanosine) was also monitored using LC/MS. Results showed that the level of oxidative DNA damage was found to be at background/endogenous levels (approximately 1-10 modified molecules/10^6 DNA bases). Nearly identical results were obtained when measuring the nucleosides with LC/MS. Accuracy of the measurements was achieved using stable isotope-labeled analogues of modified DNA bases as internal standards. The results show that the obtained tissue-engineered skin did not contain any elevated levels of oxidative DNA damage.

Impact: Fundamentals to many tissue-engineered devices are issues of inflammation associated with how biological cells respond to a given matrix or when inserted into the body. It has become increasingly clear in recent years that oxidative damage to DNA plays an important role in a number of disease processes. Oxidative damage to DNA has been implicated in cancer and several neurodegenerative diseases such as Amyotrophic Lateral Sclerosis (Lou Gehrig's disease), Alzheimer's disease and Parkinson's disease. Furthermore, the accumulation of oxidative damage to DNA especially in non-dividing cells has been postulated to be responsible for the degenerative effects of aging. Oxidative stress responses have also been implicated in apoptosis.

Future Plans: In order to assure that such composite materials are free of genetic changes that might occur from oxidative stress during the manufacturing, storage or transportation of the product, we will consider other cellular biomarkers that could be used during the development phase of tissue-engineered materials to ensure that cells have not undergone any inflammatory response during the development or shipment of the product.

Thin Films Of Collagen Effect Smooth Muscle Cell Morphology

CSTL Program: Biomaterials
Authors: *J. Elliott, and A. Tona (Brown University); J. Woodward, and P. Jones (University of Colorado Health Sciences Center); and A.L. Plant (831)*

Abstract: Collagen is the most abundant extracellular matrix protein in the body, and it is an essential component of many tissue engineered devices. In this project, we attempted to provide a reproducible method for applying collagen to surfaces on which cells can be grown, and to characterize the resulting thin films of collagen protein with respect to molecular structure and cellular response. The results showed that thin films of collagen can be reproducibly formed under conditions that produce either fibrillar or monomeric collagen, which mimic the healthy, and diseased artery, respectively. Using automated quantitative microscopic analysis we showed that the morphology and proliferation of vascular smooth muscle cells is determined by the intermolecular interactions of collagen in these thin films.

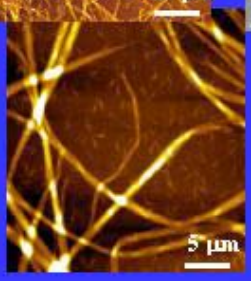
Purpose: The purpose of this study was to provide a reproducible method for applying collagen to surfaces on which cells can be grown, and to characterize the resulting thin films of collagen protein with respect to molecular structure and cellular response.

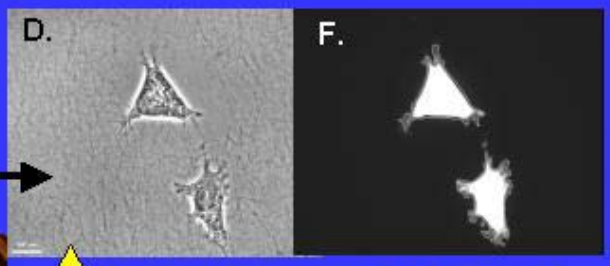
Major Accomplishments: This study showed that thin films of collagen can be reproducibly formed under conditions that produce either fibrillar or monomeric collagen, which mimic the healthy, and diseased artery, respectively. Using automated quantitative microscopic analysis we showed that the morphology and proliferation of vascular smooth muscle cells is determined by the intermolecular interactions of collagen in these thin films.

Impact: Collagen is the most abundant extracellular matrix protein in the body, and it is an essential component of many tissue engineered devices. It can assume different molecular and supramolecular structures, depending on the conditions under which it is prepared for cell culture. Standardized protocols for collagen use in cell culture studies do not exist, and rarely are the final matrices characterized. Furthermore, the cellular response to different forms of collagen is poorly studied.

Future Plans: Other rugged and reproducible thin films will be developed as essential reference surfaces on which smooth muscle and other cells can be examined for analysis of their intracellular signaling in response to their culture environment.

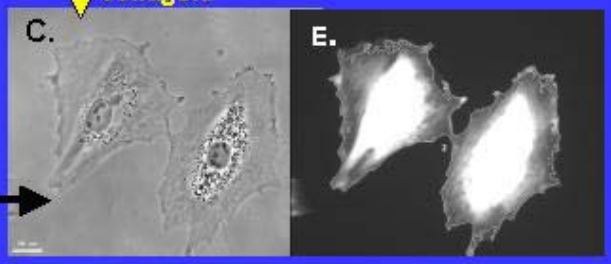
Biomimetic Surfaces of the Extracellular Matrix Protein, Collagen

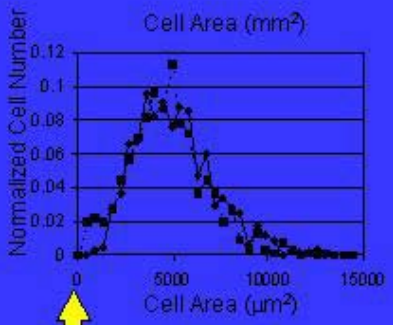




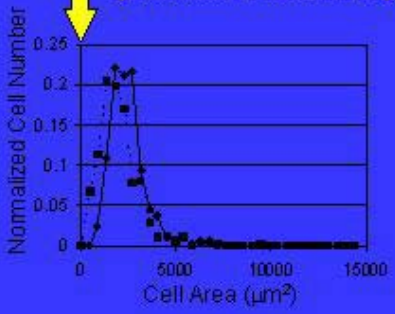
As in healthy arteries, smooth muscle cells are small and don't proliferate on thin films of collagen fibrils.

As in cells during wound healing, cells spread and proliferate on monomeric collagen.





The size of cells on thin films (solid lines) were compared to cells on thick gels (dashed lines) using automated quantitative microscopy



Y Chromosome Assays and Standards

CSTL Program: Forensics and HLS
Authors: *J.M. Butler, R. Schoske, P.M. Vallone, J.W. Redman, and M.C. Kline*

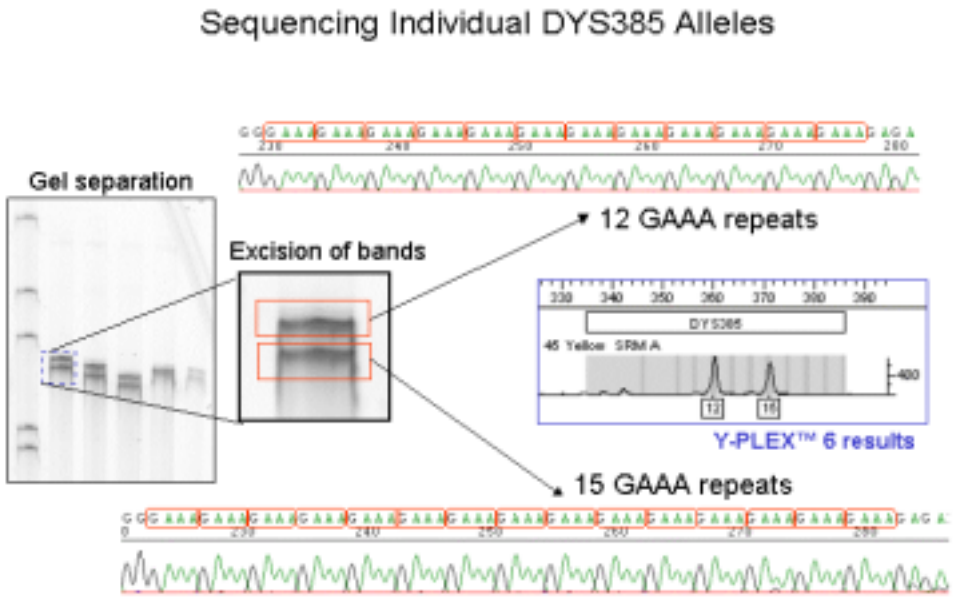
Abstract: Y chromosome short tandem repeat (STR) and single nucleotide polymorphism (SNP) markers have a number of applications in human identity testing including typing the perpetrator of sexual assault cases without differential extraction and tracing paternal lineages for missing persons investigations. In order for Y STR systems to become more widely accepted within the forensic DNA typing community, robust multiplex assays are required that will permit collection of information from many sites along the Y chromosome from a very minute amount of template DNA. We have focused on the design and development of new Y STR multiplexes as well as evaluating various markers in the same reference set of DNA samples.

Purpose: New Y chromosome assays will permit more rapid collection of information on the variability of Y STR and Y SNP markers in human populations. A standard reference material (SRM 2395) is under development to aid in reliability of measuring Y chromosome DNA typing results. This SRM contains extracted DNA from five unrelated and anonymous male individuals with sequence information from over 20 different and commonly used Y STR markers.

Major Accomplishments: NIST was asked to help organize and present keynote presentations at Y chromosome workshops held at the American Academy of Forensic Sciences (Atlanta, GA) and the International Symposium on Human Identification (Phoenix, AZ). We are part of the FBI's Scientific Working Group on DNA Analysis Methods Y chromosome subcommittee and are helping to define the core set of Y STR and Y SNP markers that will be used in future national and international DNA databases.

Impact: Promega Corporation (Madison, WI) is developing a 12plex Y STR kit that is a subset of the markers contained within the published NIST Y STR 20plex (reported in last year's technical activity report). Promega is using many of our primers and building upon the information gained in our multiplex development. Orchid Cellmark (Dallas, TX) has adopted the NIST Y STR 10plex (reported 2 years ago in a NIST technical activity report) and completed population studies and forensic validation with this multiplex assay. Orchid plans to market their kit to members of the forensic DNA typing community. Both companies are looking to NIST SRM 2395 (Human Y Chromosome Standard) to define their allele nomenclatures as is ReliaGene Technologies Inc. (New Orleans, LA), which currently has the only commercial Y STR kits available. In addition, genetic genealogy companies Relative Genetics (Salt Lake City, UT) and FamilyTree DNA (Houston, TX) are using our 20plex or portions of it in their genetic data collection. For Y SNP assays, we have beta-tested the Signet Y SNP Identification System from Marligen Biosciences, Inc. (Ijamsville, MD) and provided them with data on the usefulness of their DNA markers.

Future Plans: Completion and release of SRM 2395 to enable reliable typing of Y chromosome markers. We are in the process of examining several hundred human DNA population samples from various populations to determine the usefulness of each Y STR and Y SNP marker for distinguishing unrelated male individuals from one another.



Separation and sequencing of alleles for the Y STR marker DYS385, which is duplicated on the Y chromosome and gives rise to two bands that must be separated prior to sequencing. The number of GAAA repeats observed in the prototype SRM 2395 samples confirms allele calls by the Y-PLEX 6 kit from ReliaGene Technologies Inc. (New Orleans, LA).

Providing Assistance in the DNA Identifications of World Trade Center Disaster Victims

CSTL Program: Forensics and HLS
Authors: *J.M. Butler, P.M. Vallone, G. Spangler, and M.C. Kline*

Abstract: The high degree of fragmentation of the individuals killed in the World Trade Center (WTC) disaster of September 11, 2001 has meant that forensic DNA typing is the primary means employed to identify the victims' remains. The bones and tissue samples collected from the WTC site are highly degraded due to the severe environmental conditions experienced during the collapse of the twin towers and the fires that raged for months afterwards. The degraded samples have posed a challenge to the current state of the art in forensic DNA typing technology. Only about 50% of the specimens are yielding results with standard short tandem repeat (STR) DNA testing methods. Our laboratory at NIST is assisting the ongoing identification efforts in several different ways.

We have developed miniSTR assays that amplify smaller portions of the polymorphic sites used in standard forensic DNA tests. The smaller PCR product sizes potentially make these assays more sensitive and more capable of recovering typing information from degraded DNA specimens. Our original miniSTR assays were developed at the request of Bob Shaler who is leading the WTC DNA identification effort. A miniSTR assay dubbed "Big Mini" was provided to New York City's Office of the Chief Medical Examiner (NYC-OCME) in January 2002. Based on the groundwork laid with our miniSTRs, the Bode Technology Group (Springfield, VA) has developed their own miniSTR assays (termed "BodePlexes") that are being used to enhance the recovery of STR genotypes from WTC samples. Information from over 20,000 bones and tissue fragments is being analyzed to help make the DNA identifications of the WTC victims.

NIST has also assisted in the validation of new genetic markers that are being evaluated for their usefulness in human identification. Orchid Cellmark (Dallas, TX) has developed a panel of 70 autosomal SNP markers with PCR product sizes in the range of 60-80 bp that can produce results on degraded DNA samples. In a matter of a few weeks, the NIST team developed new assays with the Orchid markers and used them to confirm the Orchid SNP typing results. Over 500 SNP calls were evaluated and found to be fully concordant between the two different assays giving greater confidence that the Orchid results are accurate.

In addition, John Butler serves on the WTC Kinship and Data Analysis Panel (KADAP) organized by the National Institute of Justice. The KADAP panel is composed of 25 experts from around the country to evaluate new technologies and genetic analysis approaches that might be used on the WTC samples.

Major Accomplishments: Several presentations have been made on the miniSTR assays to the FBI's Scientific Working Group on DNA Analysis Methods as well as the WTC KADAP panel. These miniSTRs reduce PCR product sizes by greater than 150 bp in some cases relative to commercial STR kits (Figure 1). In July 2002, NIST provided

rapid validation of allele calls for 70 autosomal SNP markers developed by Orchid Cellmark giving greater confidence in their results. New assays and approaches to analyzing these SNP markers were developed at NIST in only a two-week time frame (Figure 2).

Future Plans: Publication of lessons learned in developing these new assays for degraded DNA as well as further testing to show concordance of results and advantages over standard forensic DNA tests.

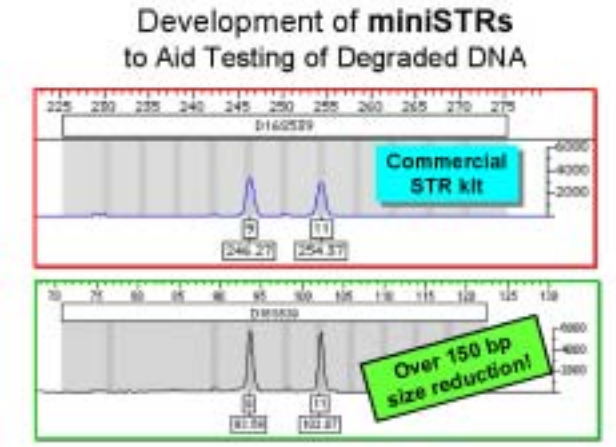


Figure 1. Comparison of allele calls at the STR marker D16S539. In both cases, a 9-11 genotype was obtained demonstrating concordance between the two assays. The new miniSTR test reduces the overall PCR product size by >150 bp.

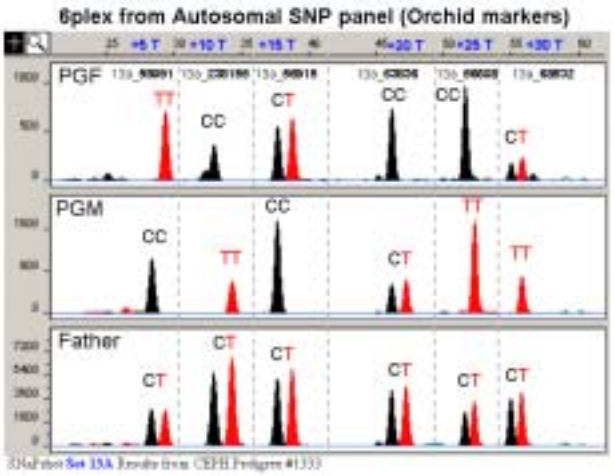


Figure 2. Results obtained from a multiplex SNP detection assay developed at NIST to evaluate the allele calls produced by the Orchid Cellmark SNP markers. A paternal grandfather (PGF), paternal grandmother (PGM), and father of family reference set were used to demonstrate expected genetic inheritance with these 6 SNP markers. The father is heterozygous at all 6 loci shown here. SNPs are being considered for use in WTC samples due to their ability to be amplified from degraded DNA as a very small PCR product (60-80 bp).

Developing and Evaluating New Forensic Tests for Probing the Mitochondrial Genome

CSTL Program: Forensics and HLS
Authors: *P.M. Vallone, M.C. Kline, and J.M. Butler*

Abstract: The mitochondrial genome consists of ~16,569 base pairs. Mitochondrial DNA is maternally inherited and can be employed in forensic investigations. The mitochondrial genome is much smaller than the nuclear human genome (3 billion base pairs). However, thousands (1000 to 5000) of copies of the mitochondrial genome are present per cell versus 2 copies of nuclear DNA. Due to the large number of mitochondrial genomes per cell it is often easier to type samples exposed to harsh environmental conditions i.e. when it is impossible to type the nuclear DNA.

A stretch of ~1100 base pairs in the control region of the mitochondrial genome is commonly analyzed for forensic purposes. The control region is highly polymorphic and is typically analyzed by DNA sequencing methods. Recently a linear array system from Roche Molecular Systems has been developed for rapid determination of common polymorphisms in the control region (specifically HV1 And HV2). Our laboratory was asked to participate in the beta testing of these "mito-strips". These linear arrays offer forensic DNA laboratories a rapid test for screening samples and excluding ones that do not match prior to the labor-intensive effort of full sequencing.

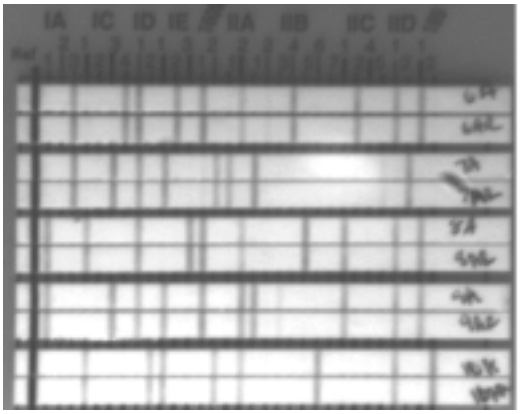
In addition to evaluating control region screening assays, we are developing new SNP assays that can probe sequence polymorphisms in the remainder of the mitochondrial genome (the coding region). This coding region test will allow for discrimination between individual who possess common (and undistinguishable) control region sequences. Currently eleven polymorphic sites are probed in our coding region multiplex test. The coding mitochondrial DNA work is being performed in collaboration with Dr. Thomas Parsons's laboratory at the Armed Forces DNA Identification Laboratory in Rockville, MD.

Purpose: The development novel multiplex SNP assays for increasing the power of mitochondrial DNA for human identifications purposes. Evaluating "mito-strip" linear arrays from Roche Molecular Systems for rapid determination of HV1 and HV2 haplogroups.

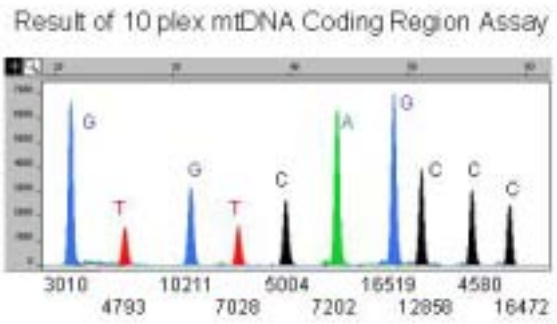
Major Accomplishments: Poster presentation describing the coding region SNP multiplexes at the 52nd Annual Meeting of the American Society of Human Genetics.

Impact: Increasing the capability/capacity of mitochondrial DNA in human identification testing. In many cases (such as mass disasters) mitochondrial DNA is the only means to identify an individual.

Future Plans: Complete testing of mito-strips on a cohort of U.S. population samples. Develop additional multiplex tests for other common HV1 and HV2 haplogroups.



Evaluation of linear array mito-strips from Roche Molecular Systems beta test.



10 mitochondrial polymorphisms are detected in a single test
using a SNaPshot assay developed at NIST

Evaluation of Extraction Methodologies for Corn kernel (Zea mays) DNA for Detection of Trace Amounts of Biotechnology-Derived DNA

CSTL Program: Food and Nutritional Products
Authors: *M.J. Holden, and J.R. Blasic, Jr.*

Abstract: Testing for the identification of biotechnology-derived plant material is currently required for compliance with labeling regulations implemented by countries other than the United States. European Union regulations mandate labeling of plant material and food as "genetically modified" if the material is present at a level of 1 % by weight. As a result, testing is important for trade in U.S. agricultural products, especially for the export of two commodity crops, soybean and corn. Testing involves detection of the specific DNA that defines the modification in the plant genome. "Needle in a haystack" is relevant imagery. The polymerase chain reaction, or PCR, is a powerful tool to detect the target DNA or "needle" by synthesizing many copies of the DNA. Specific and consistent detection is required for accurate identification of a particular genetic modification in a plant sample, especially when the modification is present in trace amounts.

The first important requirement for success with PCR is the isolation of DNA of high quality and in sufficient quantity to be representative of the sample of plant material. In this study, four investigators in two laboratories (NIST and USDA) evaluated six different methodologies. The plant material from which DNA was isolated consisted of ground corn kernels that contained 0.1 % by weight of biotechnology-derived corn of a specific type known as Star Link. DNA was extracted using five commercial kits plus a published protocol, a variation of a method known as CTAB / NaCl. The particle size of the corn flour used for the DNA isolations was evaluated along with the quality and efficiency of extraction of the DNA from ground corn. The isolated DNA was then tested in three different PCR assays to ascertain the amplifiability of the target DNA. The PCR assays resulted in some "false negative" results, that were most likely due to inhibitors of the PCR process that co-purified with the DNA.

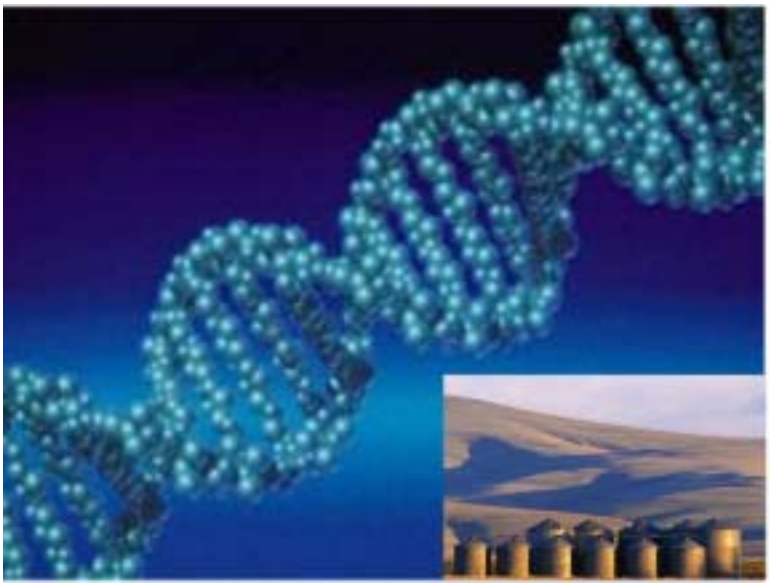
Purpose: To investigate the quality of genomic DNA, isolated using protocols based on some common principles (binding to matrix or selective precipitation), as material for testing for the presence of biotechnology-derived corn

Major Accomplishments: The study highlights several aspects that should be considered when choosing a method for DNA extraction. The sample from which DNA is extracted must be representative of the material from which it is taken. Important parameters in this case include sufficient mass and fineness of the grinding plus the extraction efficiency of the method. Many individual samples must be tested for quality and amplifiability to evaluate a particular method. While several, but not all, methods performed reasonably well, the study showed that additional purification steps are necessary to accurately and consistently detect trace amounts of target DNA. We demonstrated that a commonly used criterion for verifying the quality of isolated genomic DNA for PCR, amplification of an endogenous gene (present in all corn DNA),

is not sufficient and can lead to false negative results. Inhibitors of the PCR process, i.e. corn kernel components isolated along with the genomic DNA, are likely culprits in false negative results. Our investigations show that acidic polysaccharides act to inhibit the PCR reaction. However, starch, a major polysaccharide component of corn kernels, is not.

Impact: The results of this study are of practical importance for testing laboratories of all kinds that have a mission to identify genetic components of grain. Accurate identification is critical to US international trade in commodity crops.

Future Plans: The identification of specific inhibitors and their mode of interference in DNA amplification process should be addressed. The effect of inhibitors on the quantitation of DNA and the efficacy of various processes for removal of inhibitors will be investigated. Similar studies on DNA extraction should be conducted on relevant plant materials with different compositional analysis. It would be of value because different problems would be encountered. An example is soybean, which contains significant lipid content, not found in corn.



The Complete Mitochondrial DNA (mtDNA) Genome Sequence of Human Cell Line HL-60 and Its Inclusion in the NIST Human mtDNA Standard Reference Material - SRM 2392.

CSTL Program: Health & Medical Products/Services

Authors: *B.C. Levin[1], K.A. Holland[1,2], D.K. Hancock[1], M. Coble[3,4], T.J. Parsons[4], L.J. Kienker[5], D.W. Williams[6], G.Herrin, Jr.[6], and K.L. Richie[1]. 1. NIST, 2. Gettysburg College, 3. Georgetown University, 4. Armed Forces DNA Identification Laboratory, 5. FBI Laboratory, 6. Georgia Bureau of Investigation*

Abstract: MtDNA is used by the forensic community for human identification and by the medical community for diagnoses of human mtDNA diseases. A mtDNA Standard Reference Material (SRM 2392) prepared by NIST to provide quality control to the scientific and clinical communities when they amplify and sequence human mtDNA became available in December 1999 (1,2). This SRM includes two human DNA templates (CHR and 9947A) and all the information necessary to successfully conduct PCR amplification, cycle sequencing, gel separation, and data analysis to obtain the final DNA sequence. The sequences of 58 unique primer sets, which allow the sequencing of the entire mtDNA (16,569 bp) with no gaps, are also provided. The FBI recently requested that NIST add DNA from cell line HL-60 which has several evenly spaced polymorphisms in the mtDNA control region and no C-stretch (difficult to sequence) areas. This addition of HL-60 to the NIST SRM 2392 and an interlaboratory evaluation of the HL-60 sequence by three other laboratories have been completed. The new SRM 2392a will provide additional quality control when amplifying and sequencing human mtDNA. Corroboration of the SRM results will provide assurance that unknown DNA samples are also being amplified and sequenced correctly.

Purpose: The FBI needs DNA SRMs to provide the quality control and assurance that forensic laboratories in the U.S. are sequencing unknown DNA samples correctly. On July 15, 1998, the FBI Director signed Standard 9.5 that stated "The laboratory shall check its DNA procedures annually or whenever substantial changes are made to the protocol(s) against an appropriate and available NIST Standard Reference Material or standard traceable to a NIST standard." The FBI's Combined DNA Index System (CODIS) program now includes mtDNA sequences from unidentified remains, as well as from relatives of missing persons. In order for authorized laboratories to contribute or examine these indices, the FBI has deemed that certain quality standards must be met. In particular, a positive control from the human cell line HL-60 must be run in conjunction with their unknown samples. The primary customers are the forensic laboratories throughout the U.S. and the clinical laboratories that screen the DNA of patients for mitochondrial diseases.

Major Accomplishments: All the research at NIST and the Interlaboratory Evaluation necessary for the addition of HL-60 to SRM 2392a is completed. This research has been presented at the 12th International Symposium on Human Identification, Biloxi, MS, October, 2001; the American Academy of Forensic Sciences Annual Meeting, Atlanta, GA, February, 2002; the Third Annual Biotechnology Retreat, Shepherdstown, WV,

May, 2002; the United Mitochondrial Disease Foundation Mito Dallas 2002 Symposium and Mitochondrial Standards Workshop, Dallas, TX, June, 2002; the Third Annual DNA Grantees' Workshop, National Institute of Justice, June, 2002; and the 13th International Symposium on Human Identification, Phoenix, AZ, October, 2002. A paper has been written and has been approved by WERB (3). A report has been written to the National Institute of Justice.

Impact: The inclusion of HL-60 DNA in SRM 2392a will enhance this standard reference material's utility to all the U.S. forensic laboratories that fall under the jurisdiction of the FBI. Since we are providing all the data on the entire mtDNA (all 16,569 bp), it will also provide another positive control for the clinical laboratories that screen patients for mtDNA mutations and diseases and for the toxicologists who screen mtDNA for mutations. It also provides a positive control for any investigator who wants assurance that their laboratory is amplifying and sequencing DNA (nuclear or mitochondrial) correctly.

Future Plans: The HL-60 DNA needs to be added to SRM 2392a. The report of analysis and certificate of analysis need to be written and the SRM will be ready for sale.

References:

1. Levin, B.C., Cheng, H., Reeder, D.J. 1999. A human mitochondrial DNA Standard Reference Material for quality control in forensic identification, medical diagnosis, and mutation detection. Genomics 55:135-146.
2. Levin, B.C., Cheng, H., Kline, M.C., Redman, J.W., Richie, K.L. 2001. A review of the DNA standard reference materials developed by the National Institute of Standards and Technology. Fresenius J. Anal. Chem. 370:213-219.
3. Levin, B.C., Holland, K. A., Hancock, D.K., Coble, M., Parsons, T.J., Kienker, L.J., Williams, D.W., Herrin, Jr., G., and Richie, K.L. The Complete Mitochondrial DNA (mtDNA) Genome Sequence of Human Cell Line HL-60 and Its Inclusion in the NIST Human mtDNA Standard Reference Material - SRM 2392. (Through WERB Review).

Mitochondrial DNA Mutations in Patients with Myelodysplastic Syndromes

CSTL Program: Health & Medical Products/Services
Authors: *M.G. Shin, S. Kajigaya, and N.S. Young (Hematology Branch, National Heart, Lung, and Blood Institute, National Institutes of Health); and B.C. Levin (831)*

Abstract: Recently, a number of papers have been published that equate mtDNA mutations with specific diseases, including cancer, which had previously not been associated with mtDNA mutations. The functional relevance of these mutations (i.e., whether they have a causal relationship or could serve as biomarkers for early detection of the disease) has not been proven. In this research, we examined the mtDNA mutations found in patients with Myelodysplastic Syndrome by amplification and sequencing of the entire human mtDNA (16569 bp) and compared the results with a comparable normal control group. Myelodysplastic Syndrome is defined as "any of a group of related bone marrow disorders of varying duration preceding the development of overt acute myelogenous leukemia; they are characterized by abnormal hematopoietic stem cells, anemia, neutropenia, and thrombocytopenia. Splenomegaly, hepatomegaly and lymphadenopathy may not occur until the onset, often explosive, of leukemia" (Dorland's Illustrated Medical Dictionary, edition 28). Our results do not support a role for mitochondrial genomic instability in myelodysplasia and they fail to confirm previous reports of significant or widespread mitochondrial mutations in this disease.

Purpose: To try to confirm previously described mitochondrial DNA mutations in sideroblastic anemia, and "hot spots" in the Cytochrome C Oxidase I and II genes by using the gold standard of PCR and sequencing the entire mitochondrial DNA genome of both patients with Myelodysplastic Syndrome and a control group without the disease.

Major Accomplishments: The entire mtDNA genome (16,569 bp) in 10 patients with Myelodysplastic Syndrome and 8 normal controls was amplified by PCR and sequenced. The sequences were analyzed and compared to each other as well as recognized mitochondrial DNA databases. Overall, there was no increase in the number of mtDNA genes harboring polymorphisms or "new" mutations between our patients and normal controls, although there were a few more mtDNA changes resulting in amino acid changes in myelodysplasia (16 in 10 patients versus 9 in 8 controls). Thirty new mutations, all nucleotide substitutions, were found among the ten patients, distributed throughout the mitochondrial genome; five mutations resulted in amino acid changes, but none in the controls. We were not able to confirm previously described mutations in sideroblastic anemia, nor "hot spots" in the Cytochrome C Oxidase I and II genes. Our data do not support a major role for mitochondrial genomic instability in Myelodysplasia, and they fail to confirm previous reports of significant or widespread mitochondrial mutations in this disease. Modest changes in mutation numbers and mitochondrial microsatellites may be evidence of increased mutagenesis in mtDNA, or, more likely, a reflection of limited clonality among hematopoietic stem cells in this bone marrow failure syndrome. A paper has been written and accepted by the journal Blood and is in press (1).

Impact: To date, over 100 point mutations and more than 200 deletions and rearrangements in mtDNA have been associated with disease, and new mutations are being described every year (2). Ironically, the database of "normal" mtDNA sequences is relatively limited. The classic Cambridge Reference Sequence, based on a consensus analysis of a placenta, the HeLa cell line and some information from the bovine sequence, has recently been corrected based on a reanalysis of the original placenta (3). While other sequences have been reported in the literature and to computerized databases, the origin of the tissues tested has often been from individuals suspected of harboring pathologic mutations or their family members. Even the distinction between polymorphisms, which are common, and new mutations, is poorly demarcated. For these reasons, in the current study, we also undertook to determine the sequence of bone marrow mtDNA from a comparable number of age-matched normal volunteers. This work points out the importance of preparing a database of the sequences of the entire mtDNA of normal individuals such as those that are available in the NIST SRM 2392 (4). It also indicates the importance of determining the functional relevance of mtDNA polymorphisms and mutations and that without the proper controls, the importance of these differences may be overestimated.

Future Plans: We plan to continue our collaboration with the National Heart, Lung and Blood Institute at NIH, however, this particular project is finished, the manuscript is written, through review, accepted by the journal Blood and is in press.

References:

1. M.G. Shin, S. Kajigaya, B.C. Levin, N.S. Young. Mitochondrial DNA Mutations in Patients with Myelodysplastic Syndromes. Blood (in press).
2. R.K. Naviaux (2000) Mitochondrial DNA disorders. Eur J Pediatr. 159 Suppl 3:S219-S226.
3. R.M. Andrews, I. Kubacka, P.F. Chinnery, R.N. Lightowlers, D.M. Turnbull, N. Howell.1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet. 23:147.
4. B.C. Levin, H. Cheng, D.J. Reeder. 1999. A human mitochondrial DNA standard reference material for quality control in forensic identification, medical diagnosis, and mutation detection. Genomics 55:135-146.

Design and Use of a Peptide Nucleic Acid for the Detection of the Heteroplasmic Low-Frequency MELAS (Mitochondrial Encephalomyopathy, Lactic Acidosis and Stroke-Like Episodes) Mutation in Human Mitochondrial DNA

CSTL Program: Health & Medical Products/Services
Authors: *D.K. Hancock, F.P. Schwarz, and B.C. Levin (831); F. Song (U.Md. CARB); and L-J.C. Wong (Georgetown University Medical Center)*

Abstract: A multitude of mitochondrial DNA (mtDNA) diseases have been correlated with single nucleotide polymorphisms (SNPs), mutations, insertions, and deletions. Most of these diseases are neuromuscular, but deafness, diabetes, epilepsy, progressive dementia, hypoventilation, cardiac insufficiency, renal dysfunction, and sudden onset blindness are some of the other symptoms of mtDNA mutations. Most pathogenic human mtDNA mutations are heteroplasmic (i.e., the mutant mtDNA coexists with the normal mtDNA in the tissue or cell). The mutation level varies with the tissue and is often difficult to detect (especially in blood samples) when very low levels of the mutant exist in a population of wild-type mtDNA molecules. We are using a simple methodology to detect low levels of the single base pair heteroplasmic MELAS (Mitochondrial Encephalomyopathy, Lactic Acidosis and Stroke-like Episodes) (A3243G) mutation. A series of peptide nucleic acids (PNAs) was designed to bind to the wild-type mtDNA, thus reducing the polymerase chain reaction (PCR) amplification of the wild-type mtDNA to background levels while permitting the mutant DNA to become the dominant product and readily discernable. This methodology permits detection of the MELAS A3243G mutation in asymptomatic or symptomatic carriers with low to undetectable blood levels of this mutation. Any laboratory with PCR and sequencing capabilities will be able to use this methodology to detect this mutation at levels as low as 0.1% and verify the presence of the disease.

Purpose: This work addresses the problem of detecting low-frequency (below 20%) single mtDNA mutations in order to help in the diagnosis of mitochondrial diseases and provide more predictive genetic counseling and assistance in screening of *in vitro* fertilized embryos to prevent the inheritance of these devastating mitochondrial diseases. Twenty percent is about the lowest concentration that is detectable by PCR and sequencing which is considered the gold standard for detection of mutations. Our technique could also be useful in the screening of mixed population samples for low-frequency SNPs. Now that the human genome has been sequenced, the next big effort is to detect the SNPs and mutations, discover their effects (diversity or the cause of diseases) and determine their population frequency.

Major Accomplishments: We designed, synthesized, characterized and purified a series of PNAs and tested them to determine the best one and the optimal conditions for blocking the amplification of the wild-type mtDNA while allowing the mutant DNA to amplify. We determined the optimal concentration of the PNA to be about 2 μmol/L. We examined eight MELAS patients as well as control healthy individuals. In all the MELAS patients, we were able to show that mutations that were not visible or barely visible in the

sequence electropherogram in the absence of PNA became the dominant peak and readily detectable in the presence of our designer PNA. The lowest detectable level of the mutation was 0.1%. Portions of this research have been presented at the Biotechnology Offsite(April 2001), the 12th International Meeting on Human Identification (October 2001), the Third Annual Biotechnology Division Retreat (May, 2002), the Mitochondrial Standards Workshop: Laboratory Methods for the Diagnosis of Mitochondrial Disease, Dallas, Texas (June, 2002) and the Biotechnology Division Seminar (September, 2002). The manuscript on this research has been accepted by Clinical Chemistry and is in press (1).

Impact: The A3243G mtDNA point mutation is the second most common mtDNA defect. In addition to its association with MELAS, it is the most common mtDNA mutation associated with diabetes mellitus. When these results are published and when a SRM is available, there will be a considerable impact on the clinical community who screen patients for mitochondrial diseases. It will greatly facilitate the detection of a mutation that was previously extremely difficult to detect and will help to prevent false negative results. It will provide help to those doing genetic counseling and those screening human oocytes for *in vitro* fertilization processes. It will serve as a model of a methodology that scientists screening populations for specific SNPs will be able to use.

Future Plans: We plan to develop a NIST SRM for the detection of the MELAS mutation in human mtDNA. This SRM will consist of the specially designed PNA, the primers needed to amplify the region where the mutation occurs, and the optimized protocol to allow any laboratory with PCR and sequencing capabilities to use this SRM and detect the MELAS mutation in blood even if it is not detectable in the absence of the PNA. Since many of the human mtDNA diseases have been correlated with single base substitutions, we hope to add more specific PNAs and primers to this SRM in the future to detect additional human mtDNA mutations. We also plan to extend this work to study the prevalence of the A3243G mtDNA mutation in diabetes mellitus patients.

Reference:

1. D.K. Hancock, F.P. Schwarz, F. Song, L-J.C. Wong and B.C. Levin, Design and Use of a Peptide Nucleic Acid for the Detection of the Heteroplasmic Low-Frequency MELAS (Mitochondrial Encephalomyopathy, Lactic Acidosis And Stroke-Like Episodes) Mutation in Human Mitochondrial DNA. Clinical Chemistry (in press).

Evaluation of Genotyping Technologies for Estimating Single Nucleotide Polymorphism (SNP) Allele Frequencies in Pooled Samples

CSTL Program: Health & Medical Products/Services
Authors: *B.C. Levin, D.K. Hancock, and K.L. Richie (831); J. Schloss (National Human Genome Research Institute, NIH); and D. Bartley (Center for Medical Genetics, Johns Hopkins University School of Medicine)*

Abstract: Now that the human genome has essentially been completed, two of the next major and daunting tasks are to detect all the SNPs that are present and to understand their role in contributing to diseases. It is estimated that 10-30 million SNPs exist in the human population. The contribution of any particular SNP to disease is typically determined in association studies that seek to establish a correlation between the presence of certain alleles and the occurrence of the disease. Common, complex diseases are particularly challenging since they have multiple genetic determinants each of which plays a fractional role in its etiology. Therefore, detecting the contribution of each of several genes requires screening of large numbers (usually thousands) of individuals, each of whom must be genotyped for thousands to tens of thousands of SNPs. With current technology, this is a difficult, time-consuming, and costly endeavor. In this research, we are examining SNPs in pooled samples of individuals with and without specific diseases to determine the best screening methods to detect SNPs, especially those that may be present in low concentrations in these pooled samples. This research has just started and only preliminary experiments have been conducted. The sensitivity of the various methods for detecting allele frequency differences between pooled disease populations and controls will be determined. A consortium of laboratories with expertise in different methods will be formed. Initially, however, the samples containing the SNPs, which will be pooled and will be sent to the members of the consortium, need to be identified, obtained, analyzed to ensure the SNPs are present, and the pooled samples need to be constructed and tested. The determination of the best techniques to screen for disease-related SNPs in pooled samples will enable the healthcare community to detect the individuals with specific diseases in a timely and cost-effective manner.

Purpose: The objective of this research is to evaluate the currently available state-of-the-art methodologies to determine SNP allele frequencies in pooled samples generated from individuals that have the disease, on the one hand, versus from individuals that do not have the disease. The problem is how to determine the best disease diagnostic techniques without having to screen each individual. The primary customers will be the healthcare community and the patients for whom they care.

Major Accomplishments: We have identified samples containing homozygous SNPs and their counterparts without the SNPs. We now have those samples plus the primers needed to amplify the areas of interest. We tried to amplify all of the samples, but one set (with and without the SNP) did not amplify and one set produced double bands. All the amplified samples were sequenced on the ABI 310 to ensure the SNPs are present and that data is currently being analyzed.

Impact: Now that the human genomes of a few individuals have been deciphered, we need to detect all the SNPs that are present in various individuals and groups of people and to understand their roles in contributing to diversity and diseases. The impact of this approach will be to reduce the time and cost of examining each individual by screening pooled samples. This research will benefit the healthcare community and the affected patients.

Future Plans: 1. The currently available state-of-the-art genotyping technologies need to be tested to decide those that are best suited for this task, cost-effective, and produce results in a timely manner; 2. Laboratories with expertise in various SNP detection methodologies need to be identified and invited to join a consortium to examine and compare the various platforms; 3. Once we are assured that the SNPs are present in our samples, we need to prepare pooled mixtures to provide the appropriate allele frequencies and test them again to determine if the mixtures were prepared correctly; 4. The coded samples will be distributed to the consortium members, each of whom will use their own techniques to determine the location of the SNPs and to quantitate the allele frequencies; and 5. A manuscript will be prepared for publication.

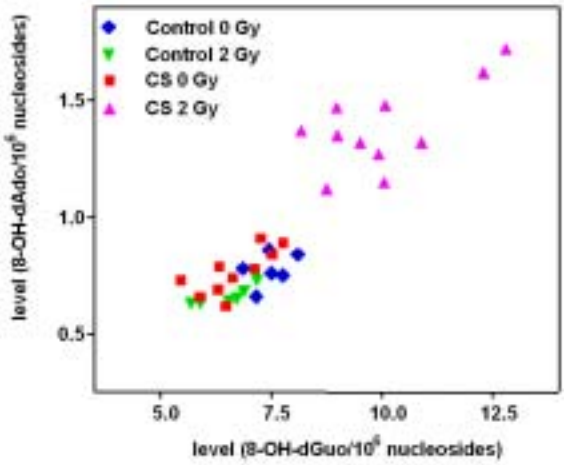
The Role of the Gene of Cockayne Syndrome in Cellular Repair of Oxidative DNA Damage

CSTL Program: Health & Medical Products/Services
Authors: *M. Dizdar, P. Jaruga, and H. Rodriguez (831); and J. Tuo, and V. Bohr (National Institute on Aging, NIH)*

Abstract: Cockayne Syndrome (CS) is a human genetic disorder with diverse clinical symptoms that include hypersensitivity to sunlight, severe mental and physical growth retardation, progressive neurological and retinal degeneration, and skeletal abnormalities. CS has two complementation groups, CS-A and CS-B. The gene of CS-B (*CSB* gene) encodes a protein (CSB protein), which is known to play a role in the cellular repair of DNA damage, but it may participate in other pathways of DNA metabolism. In this study, the role of the CSB protein in the cellular repair of oxidative DNA damage was investigated. The results of our study suggest that the accumulation of oxidative stress-induced DNA lesions in genomic DNA of the CSB patients may contribute significantly to the pathogenesis of the human disorder CS.

Purpose: Cockayne Syndrome (CS) has two complementation groups, CS-A and CS-B. The gene of CS-B (*CSB* gene) encodes a protein (CSB protein), which is known to play a role in the cellular repair of DNA damage, but it may participate in other pathways of DNA metabolism. In this study, the role of the CSB protein in the cellular repair of oxidative DNA damage was investigated.

Major Accomplishments: Transformed human cell lines with site-directed mutations in the *CSB* gene were established. These cell lines were then used to study phenotypical changes affected by the mutations, including DNA repair by whole cell extracts and the accumulation of oxidative DNA damage in genomic DNA of cells after exposure to oxidative stress. Cells were exposed to low doses of ionizing radiation such as 2 Gray to cause oxidative stress. It was found that the mutant cells and those with the deleted *CSB* gene had greater sensitivity than wild-type cells to ionizing radiation. The mutant cells had lower activity of DNA repair as shown by determination of activities of DNA repair enzymes. The results showed that the biological functions of the CSB protein in DNA repair might be mediated by distinct functional motifs of the protein. In addition, the accumulation of two well-known products of oxidative DNA damage, i.e., 8-hydroxy-2′-deoxyguanosine (8-OH-dGuo) and 8-hydroxy-2′-deoxyadenosine (8-OH-dAdo) in genomic DNA of cells was measured using the technique of liquid chromatography-mass spectrometry. These



Accumulation of 8-OH-dGuo and 8-OH-dAdo in genomic DNA of Cockayne syndrome patients after exposure to γ-radiation

products accumulated to a greater extent in mutant cells and in cells with the deleted *CSB* gene than in wild-type cells after exposure of cells to ionizing radiation. This work was then extended to primary fibroblasts from eleven CS patients and six control individuals. These cells were exposed to 2 Gray of ionizing radiation to induce oxidative DNA damage and were then allowed to repair the damage. No significant change in background levels of the afore-mentioned compounds was observed in DNA of fibroblasts of normal individuals, indicating complete cellular repair. In contrast, cells from CS patients accumulated significant amounts of these lesions, providing evidence for a lack of DNA repair. Taken together, this study suggests that the accumulation of oxidative stress-induced DNA lesions in genomic DNA of the CSB patients may contribute significantly to the pathogenesis of the human disorder CS.

Impact: The role of the CSB protein in the cellular repair of oxidative DNA damage was investigated. This type of DNA damage results in living cells from normal cellular metabolism and from oxidative stress caused by DNA-damaging agents such as UV- and ionizing radiations, and carcinogenic compounds. Transformed human cell lines with site-directed mutations in the *CSB* gene were established. Our study suggests that the accumulation of oxidative stress-induced DNA lesions in genomic DNA of the CSB patients may contribute significantly to the pathogenesis of the human disorder CS.

Cockayne syndrome sufferers have multi-systemic disorders due to a defect in the ability of cells to repair DNA that is being transcribed. (Lehman, A.R. Trends Biochem. Sci. 20, 402-405, 1995)

Future Plans: The experimental drug tirapazamine, is lethal only to cells with very low concentrations of oxygen--a characteristic of many tumor cells in most types of human cancers. Because these hypoxic tumor cells tend to multiply more slowly than others, they are resistant to the most commonly used cancer drugs, which target cells that actively multiply. Hypoxic tumor cells are also highly resistant to radiation. Tirapazamine is the first drug to take advantage of hypoxia and shows great clinical promise (currently being examined in several phase III studies). It is well established that DNA is an important cellular target for tirapazamine; however, the structural nature of the DNA damage inflicted by this drug remains poorly understood. As part of an effort to understand the chemical events responsible for the hypoxia-selective cytotoxicity of tirapazamine, we will conduct studies to characterize this drug's ability to damage DNA.

Discovery of A Critical DNA Repair Enzyme

CSTL Program: Health & Medical Products/Services
Authors: *M. Dizdar, and P. Jaruga (831); and T. Hazra, and S. Mitra (University of Texas Medical Branch, Galveston, TX)*

Abstract: A novel human DNA repair enzyme that is involved in the repair of oxidative DNA damage has been discovered. Mass spectrometric measurements determined the specificity of this enzyme and provided evidence that this enzyme is a DNA glycosylase and removes two biologically significant lesions from DNA. DNA glycosylases are repair enzymes that are involved in the first step of a complex DNA repair process in cells and remove modified bases from DNA. This enzyme named NEH1 possesses two characteristics. First, it works on critical portions of human DNA that are actually active. Second, when DNA is replicating, cells produce more NEH1 as if in an effort to repair errors caused by oxidative DNA damage before they are fixed in the next generation causing mutations. The coupling of repair to DNA replication indicated that the repair of active genes is different from the repair of the bulk of the genome. The paper that describes these results was featured on the cover page of the *Proceedings of the National Academy of Sciences*.

Purpose: The identification and characterization of DNA repair enzymes might be the first step toward understanding as to how oxidative stress produces health problems and/or exacerbates the existing diseases. This study was conducted to determine the specificity of one important human DNA repair enzyme, NEH1 using mass spectrometric techniques.

Major Accomplishments: A novel human DNA repair enzyme that is involved in the repair of oxidative DNA damage has been discovered. In the human genome database, sequences were identified that are similar to those known to code for DNA repair enzymes in the bacterium *Escherichia coli*. Following identification of the sequences, a previously unknown protein was produced. This protein was characterized by various techniques. Mass spectrometric measurements proved this enzyme to function as a DNA glycosylase when tested on oxidatively damaged DNA *in vitro*, and to remove two biologically important lesions from DNA. DNA glycosylases are repair enzymes that are involved in the first step of a complex DNA repair process in cells and remove modified bases from DNA. This glycosylase named NEH1 possesses two characteristics. First, it works on critical portions of human DNA that are actually active. Second, when DNA is replicating, cells produce more NEH1 as if in an effort to repair errors caused by oxidative DNA damage before they are fixed in the next generation causing mutations. This coupling of repair to DNA replication indicates that the repair of active genes is different from the repair of the bulk of the genome. NEH1 was found to be highest in the liver, pancreas and thymus. The paper that describes these results was published in *Proceedings of the National Academy of Sciences* and featured on the cover page of this journal.

Impact: Living organisms are constantly exposed to damaging agents that cause damage to their DNA. Oxidative DNA damage results from normal cellular metabolism and oxidative stress, and is implicated in numerous human diseases from cancer to neurodegenerative diseases such as Alzheimer's disease to the normal aging process. Repair of DNA damage in cells is one of the essential events in all life forms. Detailed knowledge of mechanisms of DNA damage and repair might lead modulation of DNA repair. This in turn might lead to drug developments and clinical applications including the improvement of cancer therapy by inhibiting DNA repair in drug- or radiation-resistant tumors and/or the increase in the resistance of normal cells to DNA damage by overexpressing DNA repair genes.

Future Plans: A study of repair of oxidative DNA base damage by mouse NEIL1 protein in collaboration with State University of New York at Stony Brook will be performed.

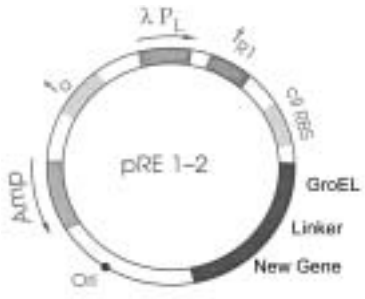
Production of Soluble And Enzymatically Active Gene Products (Proteins) in Escherichia Coli.

CSTL Program: Pharmaceuticals and Biomanufacturing
Author: *P. Reddy*

Abstract: In this project, the recombinant expression plasmid (pRE) vector, which has tightly controlled transcription / translation signals and was originally developed for cloning and overproduction of lethal proteins, was fused with a segment of the gene for a chaperone protein GroEL which is a universal protein folding machinery in the cell. A test gene (M. smegmatis calmodulin) was chosen for making fusion to GroEL in the newly constructed plasmid vector. The test gene product was not expressed at all by itself perhaps due to proteolytic cleavage in the cell. Upon expression of the GroEL:calmodulin gene fusion in E. coli, we observed that calmodulin gene was successfully expressed as a fusion protein. We subsequently purified the fusion protein, cleaved the fusion protein with enterokinase and released calmodulin from GroEL. Hence, in principle we have constructed a plasmid vector for expression of soluble proteins.

Purpose: In the protein expression endeavors, we frequently experience the production of inactive and insoluble protein aggregates called "inclusion bodies" or no protein expression at all due to proteolytic degradation. The framework of this project is directed towards engineering a universal protein expression plasmid vector that we hope to provide a solution to these frequently encountered problem in protein expression.

Major Accomplishments: We used the recombinant expression plasmid (pRE) vector with tightly controlled transcription / translation signals, which was originally developed for cloning and overproduction of lethal proteins. We then fused a segment of the gene for a chaperone protein GroEL which is a universal protein folding machinery in the cell. A short DNA molecule that codes for a peptide linker with a specific protease (enterokinase) recognition sequence for subsequent release of the protein and a unique restriction endonuclease recognition sequence (CATATG) were incorporated downstream of the GroEL gene in the vector. All the fusions are in-frame with initiator ATG. The strategy is that the improved plasmid vector could then be used for cloning and overproduction of any gene.



One test gene (M. smegmatis calmodulin) was chosen for making fusion to GroEL in the newly constructed plasmid vector. The test gene product was not expressed at all by itself perhaps due to proteolytic cleavage in the cell. Upon expression of the GroEL:calmodulin gene fusion in E. coli, we observed that calmodulin gene was successfully expressed as a fusion protein. We subsequently purified the fusion protein, cleaved the fusion protein with enterokinase and released calmodulin from GroEL. Hence, in principle we have constructed a plasmid vector for expression of soluble proteins.

Impact: Genome sequencing of several microbes and eukaryotes have revealed wealth of knowledge on the gene segments that code for proteins. Nearly 50% of the proteins have no known function and hence termed “hypothetical proteins”. The challenges in the post genomic era are: cloning of hypothetical genes, expression of the corresponding gene product (protein), and characterization of the protein for its function. Protein expression/gene expression is of vital importance to the biotechnology/pharmaceutical industry to overproduce proteins in order to design small molecular drugs and protein pharmaceuticals and to fully understand cellular function.

Future Plans: The vector will be tested with a variety of genes that are known to produce inclusion bodies.

Rapid Analysis of the Kinetics of Enzymatic Reactions by a Novel Stopped Flow Microcalorimetry Method

CSTL Program: Pharmaceuticals and Biomanufacturing
Authors: *F.P. Schwarz (CARB/NIST); and M. Stodeman (CARB/UMBI)*

Abstract: Research in microcalorimetry has lead to the development of computer-controlled, high throughput, rapid time response and highly sensitive differential stopped flow microcalorimeters (DSFM) that are amenable to monitoring the progress of an enzyme-substrate reaction calorimetrically. DSFM data for two types of enzyme-substrate reactions were analyzed in terms of the integrated form of the Michaelis-Menten rate equation to yield values for Km, the substrate concentration required for the reaction to proceed at half its maximum velocity, and kcat, the reaction turnover number. Analysis of the isomerization of fructose-6-phosphate to glucose-6-phosphate by phosphoglucose isomerase yielded values for Km and kcat from 293.4 K to 311.5 K and values for the product-substrate equilibrium constant that agreed with literature values. Analysis of the hydrolysis reaction of four N-acetyl amino acid derivatives by acylase at 298.2 K yielded values for Km and kcat, that were in fair agreement with values obtained by other methods. The exothermic heats of reaction for each substrate-enzyme reaction were also obtained at each temperature and exhibited an increase with temperature.

Purpose: Since virtually all of the enzyme-substrate reactions involve the exchange of heat with the surroundings, the progress of any enzyme-substrate reaction can be universally and efficiently monitored by a DSFM. Rapid evaluation of enzyme kinetic parameters and their dependence on temperature and pH are needed by the biocatalyst industry and are needed to identify the function of unknown proteins.

Major Accomplishments: More than 80 % of the progress of the hydrolysis and isomerization reactions monitored by DSFM agreed with the analytical expressions derived from the intergrated form of the Michaelis-Menten rate equation. The isomerization of fructose-6-phosphate to glucose-6-phosphate by phosphoglucose isomerase exhibits significant substrate-product equilibration. This exothermic reaction was monitored at 293.4 K, 298.4 K, 303.4 K, and 311.4 K and reversibly (endothermically) at 293.4 K and 298.4 K. Values of the equilibrium constant determined from the reverse and forward reactions were in agreement with literature values. The hydrolysis of N-acetyl derivatives of L-methionine, glycine, phenylalanine, and alanine at 293.2 K by acylase is exothermic and exhibits significant acetate product inhibition. The hydrolysis of N-acetyl-L-methionine was also investigated at 288.4 K, 308.5 K, and 318.5 K. Values for Km and kcat at 293.2 K were within the wide range of literature values determined by the more conventional reciprocal velocity versus substrate concentration plots.

Output: The results of the acylase investigation were presented at the 57th Calorimetry Conference and in a publication that is in press in *Analytical Biochemistry.* A second

paper on the acylase reaction was submitted to WERB and a third one on the isomerization reaction is in preparation.

Future Plans: Plans are underway to examine other types of enzyme reactions by DSFM.

Measuring Structural Changes in G-protein Peptides Upon Binding a Soluble Mimic of Activated Rhodopsin: Development of an NMR-based Drug Screening Approach for GPCRs

CSTL Program: Pharmaceuticals and Biomanufacturing
Authors: *K.D. Ridge, and J.P. Marino (831); N.G. Abdulaev (CARB/UMBI); and D.M. Brabazon (Loyola College in MD)*

Abstract: Although a crystal structure for the integral membrane G-protein coupled receptor (GPCR) rhodopsin is now available, portions of the cytoplasmic surface are not well resolved and the structural basis for its interaction with the G-protein (G_t) is unknown. Previous studies using soluble mimics of light-activated rhodopsin have shown that grafting defined segments from the cytoplasmic region onto a surface loop of thioredoxin is sufficient to confer G_t activation. To assess whether these mimics can also provide structural insights into the interaction between light-activated rhodopsin and G_t, the ability of a thioredoxin fusion protein comprised of rhodopsin's second and third cytoplasmic loops to bind G_t α-subunit ($G_{t\alpha}$) peptides was examined by NMR spectroscopy. Titration experiments show that peptides corresponding to the carboxyl-terminus of $G_{t\alpha}$ bind to this soluble mimic and undergo small, but significant structural changes in the bound state. Remarkably, the peptides adopt a C-cap like structure similar to that observed upon the binding of $G_{t\alpha}$ peptides to intact, light-activated rhodopsin. These findings suggest that this functional mimic of light-activated rhodopsin is also a structural mimic for the signaling state and provides a novel solution–based structural assay for screening small molecule inhibitors that potentially disrupt activated GPCR/G-protein interactions.

Purpose: A comprehensive understanding of the molecular details governing activated GPCR/G-protein interactions is important to the development of drugs that alleviate a number of human diseases. However, the lack of high-resolution structural data for this pharmacologically tractable class of integral membrane protein receptors remains a major gap in our knowledge. We have been developing and evaluating alternative approaches to obtain approximate structural information for GPCR's by designing and characterizing soluble mimics of rhodopsin functions. It is anticipated that such approaches will facilitate the structural analysis of receptor/ligand and receptor/G-protein interactions.

Major Accomplishments: NMR studies show that a soluble mimic for the activated state of rhodopsin binds to $G_{t\alpha}$ carboxyl-terminal peptides and induces structural changes (formation of a C-cap structure) akin to those previously observed upon binding of $G_{t\alpha}$ peptides to native, light-activated rhodopsin. These findings indicate that segments from the cytoplasmic surface of rhodopsin can be assembled on a soluble protein scaffold to elicit functionally important structural changes in a cognate G-protein.

Impact: This study represents the first example of an engineered soluble protein that appears to function through a structural mechanism that is analogous to what has been observed for an intact, activated GPCR (light-activated rhodopsin). The approaches

developed here offer a promising strategy for "solubilizing" the functions of other GPCR's in order to expedite a structural understanding of G-protein binding and specificity and to the application of drug discovery efforts focused on this interaction.

Future Plans: Since this approach provides a relatively simple solution–based structural assay for the interaction of an "activated" GPCR with its cognate G-protein, future efforts will be devoted to applying these methods for screening small molecule inhibitors that potentially disrupt the rhodopsin/G_t interaction. This will be accomplished using fluorinated derivatives of the $G_{t\alpha}$ peptides that will allow the use of ^{19}F-NMR methods to monitor the formation of the C-cap structure (a molecular signature) in the presence of a wide variety of small molecule compounds.

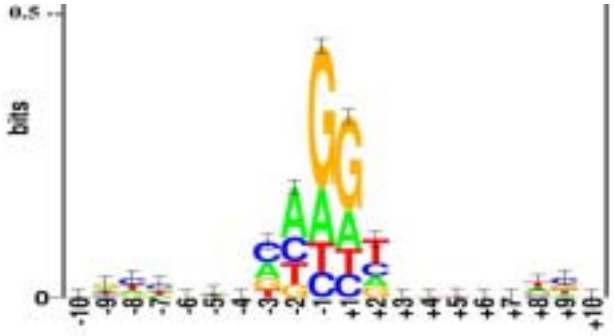
Target Site of Intron Gain Inferred by a System for Phyloinformatic Analysis (SPAN)

CSTL Program: Pharmaceuticals and Biomanufacturing
Authors: *W-G. Qiu, and A. Stoltzfus (Center for Advanced Research in Biotechnology ((CARB)), NIST); and N. Schisler (Pomona College)*

Abstract: Our goal is to develop a software system that will allow evolutionary analyses of sequence features to be carried out using hundreds of sequence families at a time. Such a System for Phyloinformatic Analysis (SPAN) will provide the bioinformatics community with a rigorous high-throughput method of comparative sequence analysis. The test-bed for development of SPAN is a set of problems regarding the evolution of introns that have withstood a decade of analysis using more conventional methods. On the hypothesis that introns are gained non-randomly, we aim to infer a precise intron "target site", and then use this to predict various aspects of the distribution of introns. In place of the expert judgement that would be used conventionally to tailor an evolutionary analysis for each sequence family, we use automated methods that assess the reliability of information, so as to allow an objective "high-grading" approach in which conclusions are based on only the most reliable parts of the analysis. Based on this kind of analysis applied to a set of 10 sequence families, we find clear evidence that introns are gained non-randomly at sequences favoring the nucleotide CAG^GT.

Purpose: Due to its proven utility, comparative analysis of sequences— comparing homologous genes, proteins or genomes to derive clues about structure and function— plays a central role in many areas of biotechnology (genome annotation, drug target discovery, biomolecule engineering, medical genetics). The most rigorous methods of comparative analysis are evolutionary ones, based on identification and alignment of homologous sequences, inference of a phylogenetic tree, and phylogenetic reconstruction of changes associated with biologically relevant differences. Due to its complexity and its reliance on human supervision, this type of analysis conventionally is done by an expert, one sequence family at a time. Our goal is to develop a system that will allow rigorous analyses of sequence features to be carried out using hundreds of sequence families at a time. As a test-bed for this System for Phyloinformatic Analysis (SPAN), we are tackling a set of problems regarding the evolution of introns that have withstood a decade of analysis using more conventional methods. On the hypothesis that introns are gained non-randomly, we aim to develop a precise characterization of an intron "target site", and then use this model to predict various non-random aspects of the distribution of introns.

Major Accomplishments: Data from ten sequence families comprising 677 genes were collated, the sequences were multiply aligned, and family trees were inferred. The resulting data sets were loaded into the SPAN



database and augmented with probabilities of ancestral intron states estimated using a Bayesian method developed at CARB. These procedures are not fully automated as yet. Database tools were developed to identify patterns of intron gain. In place of the expert judgement that might normally be used in this kind of analysis to tweak alignments and trees, and to ignore dubious data points, our system allows us to implement an objective "high-grading" approach based on explicit measures of the reliability of alignment columns, phylogenetic tree branches, and reconstructed events of intron gain and loss. The resulting intron "target" pattern shown here is based on the most trustworthy parts of a large set of data. Preliminary results suggest that this pattern can account quantitatively for patterns in the distribution of intron phases that previously were poorly understood.

Impact: At the very least, this system will be useful in deciphering genome organization, since it will lead to a much greater understanding of one of the central features of eukaryotic genome organization, namely the splitting of genes into introns and exons. The system can be adapted for the analysis of other 2-state (presence/absence) features, and in principle, for more complicated sequence features.

Future Plans: We are currently (1) testing the implications of targetted intron gain; (2) developing a more consistent and capable database interface. When the system has been scaled for hundreds of sequence families, it will be made available to the bioinformatics community.

Rapid Identification of Lead Compounds that Target Retroviral RNA and RNA-Protein Complexes using Fluorescence and NMR Spectroscopy

CSTL Program: Pharmaceuticals and Biomanufacturing
Authors: *E.S. DeJong, and J.P. Marino (831); and C. Chang, and M.K. Gilson (UMBI)*

Abstract: RNA and RNA-protein complexes can provide potentially powerful targets for the regulation of gene expression and inhibition of viral/bacterial infection. To drive RNA-based drug discovery, general approaches are required for detecting and quantifying nucleic acid-protein interactions that can be used as high-throughput screens (HTS) and for obtaining rapid structural information to guide rational drug design. Using a 2-AP fluorescence perturbation screen, we have identified small heterocyclic aromatic compounds that interact with high affinity (low μM to nM range) with the Rev Responsive Element (RRE) of HIV-1, a potential RNA target for antiviral drugs. A few of the compounds identified through the fluorescence screen, have been further shown to effectively compete with an arginine-rich peptide, derived from Rev, for binding to RRE (K_I in the μM to nM range). NMR structural analysis of one of the identified inhibitors, proflavine, bound to RRE has revealed a 2:1 (proflavine:RRE) binding stiochiometry, with the two proflavines stacked one atop of the other in the binding site. NMR experiments also demonstrate that binding of proflavine stabilizes the purine-rich high-affinity Rev binding site of RRE in a manner analogous to the native Rev-RRE interaction.

Purpose: The development of specific inhibitors of protein-RNA complexes is of significant interest to the pharmaceutical industry because these complexes provide potentially powerful targets to regulate gene expression and inhibit bacterial/viral infection. For instance, the inhibition of proteins that are involved in mRNA processing, transcription, or retroviral replication could provide powerful new drug targets to combat viral infections, fight cancer, or enhance the effectiveness of existing chemotherapeutic agents. A key roadblock to the realization of these goals is the availability of a rapid and sensitive assay to measure and quantify the binding of proteins and ligands to RNA. The goal of this research is to explore and develop NMR and fluorescence-based methods suitable for screening compound libraries for inhibitors that block these protein-RNA interactions.

Major Accomplishments: General NMR and fluorescence methods for rapidly screening and quantifying RNA-protein interactions have been developed and validated. Using these methods, a new class of compounds that target the HIV-1 retroviral RRE-Rev RNA-protein interaction has been identified, which now provides new lead compounds in the search for antiviral therapeutics. These compounds are found to bind as dimers and mimic the Rev peptide in their interaction with RRE.

Impact: Measurement technology developed in this study will have an impact on the biotechnology and pharmaceutical industries as well as the Structural Biology program at NIST. The research is also responsive to ATP's interest in projects directed at developing

novel approaches for manipulating protein-nucleic acid interactions for possible therapeutic benefits or medical diagnostic purposes. The development of NMR as a front-line tool in evaluating the structural basis for inhibition of protein-RNA complexes could potentially provide a paradigm for rational RNA drug design.

Future Plans: Through collaboration with researchers at the National Institute of Cancer, we plan to apply our methods to screen large public domain libraries of compounds in a search for new lead compounds against retroviral RNA targets, which will then be optimized using rational design strategies based on NMR structural analysis.

Structural and Biochemical Studies of Enzymes Along the Chorismate Pathway

CSTL Program: Pharmaceuticals and Biomanufacturing
Authors: *J.E. Ladner (831); and E. Eisenstein, and J. Parsons (CARB/UMBI)*

Abstract: Industrial production of drugs, new biopolymers and indigo dyes, can be facilitated by increased understanding of the aromatic amino acid synthesis pathway. The increased ability to manipulate this pathway through metabolic and protein engineering will save non-renewable, petroleum-based feedstock chemicals. This study provides a description of the enzymes involved in the chorismate metabolic pathway by solving their three-dimensional structures, modeling the mechanisms of the chemical transformations, and mapping pathway control nodes involved in the biocatalytic conversion of glucose to aromatic hydrocarbons. Only when the detailed three-dimensional structures are known for the enzymes can the precise enzymatic mechanisms and relationships between the structure and the physical properties be predicted. Most recently, the structure of the product of gene phzD from *Pseudomonas aeruginosa* has been solved. The phenazines are biologically active (antibacterial, antfungal, antitumor) aromatic products synthesized mainly by *Pseudomonas* and *Streptomyces* as part of their chorismate pathways. The phzD gene product is involved in the pathway in *Pseudomonas* in the production of phenazine-1-6-dicarboxylic acid.

Purpose: This work benefits pharmaceutical and chemical companies. The elucidation of natural biochemical pathways makes it easier to alter and utilize these pathways to make these and similar chemicals.

Major Accomplishments: The work this year has yielded a better understanding of the chorismate pathway. We have solved the structure of the product of gene phzD and have been able to identify the active site of the molecule.

Impact: The synthesis of the aromatic biologically active molecules is a very important area of study because these pathways do not exist in animals. They exist solely in bacteria, fungi, and higher plants. The elucidation of these pathways can provide pharmaceutical researchers with a better understanding for the amplification or inhibition of the production of these molecules.

Future Plans: We plan to continue our studies of the enzymes along the chorismate pathway. In particular, we are continuing to look at other enzymes in the phenazine pathway in *Pseudomonas* in order to more fully elucidate the mechanisms involved in the production of these biologically active products.

Theoretical Studies of Enzyme Mechanisms

CSTL Program: Pharmaceuticals and Biomanufacturing
Author: *M. Krauss*

Abstract: The reaction mechanisms of two enzymes, zinc lactamase and chorismate mutase, that are representative of their class are studied theoretically. The reaction mechanism of the di-zinc lactamase enzyme has not been determined before at the molecular level either experimentally or theoretically. An ab initio quantum mechanical and molecular mechanics (QMMM) theoretical method for calculating reaction paths in an enzyme has been applied to chorismate mutase in collaboration with scientists at the NIH. This enzyme has been the focus of many theoretical studies but the questions on details of the structure and energetics along the reaction path have persisted.

Purpose: Class B lactamases are a prime defense mechanism of pathogens against lactam anti-biotics and are of great interest medically. The structural and energetic details of the reaction path on a molecular level would be useful in the design of inhibitors. Global optimization and then ab initio dynamics on the fundamentals of enzyme reactivity are required to compare the same enzymes from different organisms to ultimately understand specificity of a specific enzyme in a given organism.

Major Accomplishments: Zinc lactamase possesses a flexible flap that forms part of the active site in binding of the substrate. This flap has to move approximately 10 Å and the theoretical solution requires a combination of standard molecular dynamics codes and chemical intuition. The reaction path for the hydrolysis of the antibiotic lactam is calculated to involve highly polarized waters connected in hydrogen bonded networks in the highly ionic active site. A range of reactive intermediates is described for the reaction. It is shown that the intermediates can form by proton transfers if the initial hydrolysis products are not removed from the active site of the enzyme by rapid solvation.

Ab initio quantum mechanical and molecular mechanics (QMMM) theoretical method for calculating reaction paths in an enzyme allows for the optimization of the entire enzyme providing the means for accurate calculation of the reaction path and also new insight into the interactions between the active site and the protein environment. One new insight is the observation that the protein environment around the enzyme active site does not change appreciably during the reaction. In another aspect of this study we have calculated the kinetic isotope effects of an enzyme reaction in the active site of the enzyme for the first time.

Impact: These theoretical approaches offer new insight into the analysis and design of the experiments. For example, measuring the chemical step is difficult and still under investigation experimentally. These calculations provide the **standard** that defines for the experimentalist when they are measuring the chemical step.

Future Plans: Theoretical dynamics studies are planned for the chemistry of adenylate cyclase and the coupling of the photoexcitation and proton pumping in ‘opsins’.

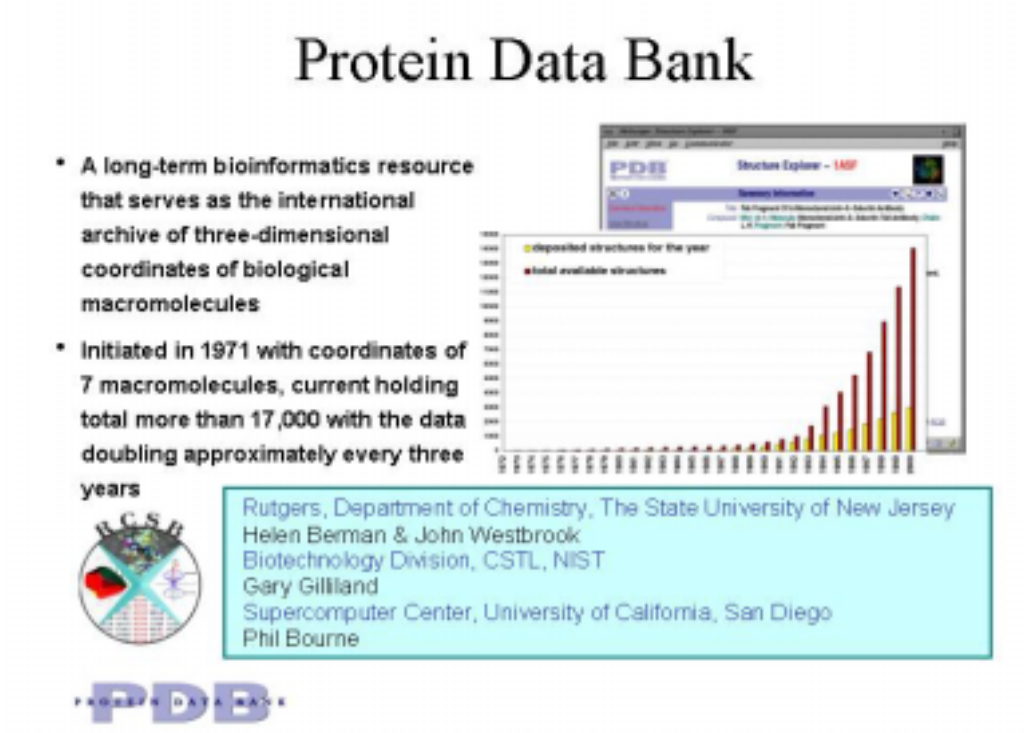
The Protein Data Bank

CSTL Program: Pharmaceuticals & Biomanufacturing
Authors: *T.N. Bhat, G.L. Gilliland, and P. Fagan*

Abstract: The Protein Data Bank (PDB) is the single worldwide repository for the processing and distribution of 3-D biological macromolecular structure data (18,294 structures in the database as of 23 July 2002). The goals of the PDB, the systems in place for data deposition and access, how to obtain further information, and plans for the future development of the resource are described in a special issue of *Acta Crystallographica D* dedicated to crystallographic databases. The article will also be published as a Chapter in a forthcoming book "Structural Bioinformatics" to be published by John Wiley & Sons (Philip E. Bourne and Helge Weissig, Editors).

Purpose: The PDB provides industry and the research community with the three-dimensional structural data of biological macromolecules that can be used for basic research and for countless applications in the pharmaceutical and biotechnology industries. During the last year, several review have been written to describe the PDB, and the data uniformity efforts of the NIST staff that lead to improved capabilities for querying the data, thereby enabling fuller utilization of the data and the resource.

Major Accomplishments: The current efforts in data uniformity and how they have been integrated into the Protein Data Bank was published in an article in Nucleic Acids Research. An invited paper in a special issue of Acta Crystallographica D dedicated to crystallographic databases has been published, and is also published as a Chapter in a forthcoming book "Structural Bioinformatics" to be published by John Wiley & Sons (Philip E. Bourne and Helge Weissig, Editors).



The article introduces and describes the goals of the PDB, the systems in place for data deposition and access, how to obtain further information, and plans for the future development of the resource. It is written to give the reader an understanding of the scope of the PDB and what is provided by the resource. It should be noted that this is a critical resource for research and educational programs with more than 100,000 hits and more than 100,000 coordinate downloads occurring each day.

NIST's Role in the PDB

Data Uniformity
- Focus on data standards to
  - Insure global uniformity
  - Facilitates data exchange between information resources
- Generate extended reference tables with standardized values for all PDB entries for use with queries
- Incorporate value-added data such as common names, synonyms, etc.

CD-ROM Distribution

Work with the NMR Community
- Adopt IUPAC nomenclature
- Development of the NMR data dictionary
- Establish "Average Structure" representation
- Geometric and experimental validation of structures

Physical Archive
- Resolve issues concerning specific entries
- Aid in uniformity and value-added annotation
- Insure disaster recovery

Physical Archive Contents
Structure Entry Files
Depositor Tapes and Disks
Archive Backup Tapes/Disks
Correspondence
Entry Data Processing Results
Newsletters
CD-ROMs
Miscellaneous Documents

Impact: The Protein Data Bank (PDB; http://www.pdb.org/) is the single worldwide archive of primary structural data of biological macromolecules. Many secondary sources of information are derived from PDB data. Structural biologists, medicinal chemists, and other biological and biomedical scientists from industry, academia, and private research efforts that are engage in protein engineering, structural genomics, rational drug design, protein stability, and other studies of biological macromolecules require information on the structures of biological macromolecules. These studies, that use the three-dimensional structural data, lead to new drugs, new proteins with desired properties for commercial applications, or new insight into biological processes. The coordinates are also essential for bioinformatics efforts since it is well recognized that three-dimensional structure of a protein is more generally conserved than is the amino acid sequence.

Future Plans: Efforts will be directed this year to enhance the query capabilities of the archive by identifying distinct files out of several million files covering several hundred GB of original data that are corrupted by redundancies.

Development of Synthetic Protein-DNA Nanostructures

CSTL Program: Technologies for Future Measurements and Standards
Author: *D.T. Gallagher*

Abstract: The design and construction of a molecular lattice is a central goal of nanotechnology because it will enable a wide assortment of applications and derivative inventions, e.g., scaffolds for attaching functional molecules such as enzymes (to make bioreactors) or electrical/optical components (to make information storage and signal processing devices). Some see the creation of such a lattice as a goal in itself, since it is consistent with the general nanotechnology trend toward finer structural control. At NIST, such lattices, by virtue of their precisely known geometry, will enable the development of powerful new measurement methods for biomolecular materials. We are exploring the biopolymers protein and DNA as materials for building the lattice. DNA is useful because it has known and predictable structural properties, forming (as duplex) long straight helical rods. Protein is necessary because DNA alone does not branch or crosslink; proteins that are already known to bind DNA will serve as lattice nodes. Of the approximately 200 DNA-binding proteins of known structure (in the Protein Data Bank), we have selected one for initial studies: a transcription factor whose geometry is potentially suitable (with some engineering modifications) for lattice design.

Purpose: To produce a bivalent DNA-binding protein module that can serve in the assembly of specific biomolecular nanostructures with measurement and nanotechnology applications. The precisely known and reproducible geometry of such constructs will enable advances in the measurement of molecular phenomena of interest to chemical-, and electronics-based industries.

Major Accomplishments: A detailed (atomic) structural model has been developed for one lattice design. This protein-DNA nanolattice has I4122 symmetry and repeat spacings of about 15 nm. The protein component is a disulfide doubled form of the dimeric (and truncated) protein component. The added cysteine residues are at the N-terminus, a position that gives the resulting dimer-of-dimers the required dihedral angle between the two bound DNA duplexes. The duplexes will then be given appropriate length (odd number of turns, divided by four) and termini (two-base sticky end overhangs) to enable assembly with other subunits. We are currently using PCR methods to assemble a synthetic gene for the engineered protein, so that protein expression and purification can proceed, followed by assembly (with separately synthesized DNA components) of the lattice-forming unit (protein-DNA complex).

Impact: None as yet as the project is still in its initial, exploratory phase.

Future Plans: The general goal of molecular lattices that can be customized to specific applications builds directly upon the present work. With appropriately designed DNA components, various constructs of increasing complexity (1-D, 2-D, 3-D, single subunit, two subunits, etc.) and a wide range of physical properties (lattice spacing, rigidity, attachment sites for functional subsystems) can be produced by small modifications in the

design. Short term plans are to (1) confirm the DNA-binding property of the cysteine-free engineered protein, (2) produce a 1-dimensional lattice by sticky-end annealing of these cysteine-free DNA-binding subunits, and measure the repeat spacings by atomic force microscopy, and (3) introduce the N-terminal cysteine and confirm disulfide-based dimerization. These steps will prepare the way for long-term plans to assemble and verify 2-D and 3-D lattices.

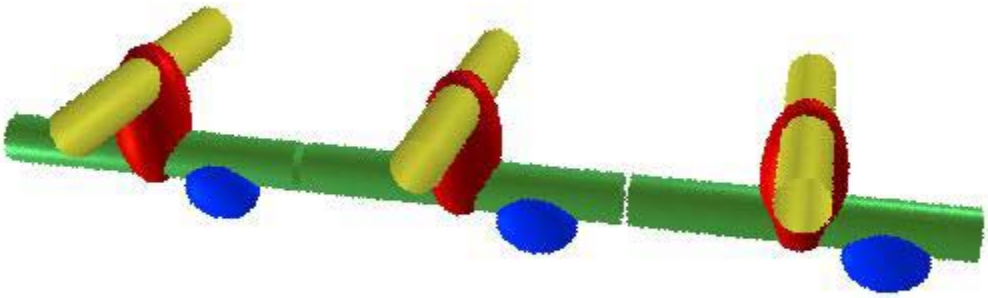


Diagram of a one-dimensional sublattice with DNA duplexes along x and y (in green and yellow), connected by proteins (red) that extend the lattice in the z direction. A functional guest molecule (blue) is shown attached to the lattice.